Long-Run Performance of Bayesian Model Averaging [1]

Adrian E. Raftery
University of Washington, Seattle

Yingye Zheng
Fred Hutchinson Cancer Research Center, Seattle

Technical Report no. 433
Department of Statistics
University of Washington

July 17, 2003

[1]Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322; email: raftery@stat.washington.edu; Web: www.stat.washington.edu/raftery. Yingye Zheng is Assistant Member, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109; e-mail: yzheng@fhcrc.org. This article is an invited discussion of the *Journal of the American Statistical Association—Theory and Methods* Invited Papers for 2003, "Frequentist Model Average Estimators," by Nils Lid Hjort and Gerda Claeskens, and "The Focussed Information Criterion," by Gerda Claeskens and Nils Lid Hjort. We are grateful to JASA—T&M Editor Frank Samaniego for inviting us to prepare it. This research was supported by NIH Grant 1R01CA094212-01 and ONR Grant N00014-01-10745. We are grateful to Merlise Clyde, Ed George, Jennifer Hoeting, David Madigan and Chris Volinsky for helpful comments.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
17 JUL 2003		**00-07-2003 to 00-07-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Long-Run Performance of Bayesian Model Averaging	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**24**	

Abstract

Hjort and Claeskens (HC) argue that statistical inference conditional on a single selected model underestimates uncertainty, and that model averaging is the way to remedy this; we strongly agree. They point out that Bayesian model averaging (BMA) has been the dominant approach to this, but argue that its performance has been inadequately studied, and propose an alternative, Frequentist Model Averaging (FMA). We point out, however, that there is a substantial literature on the performance of BMA, consisting of three main threads: general theoretical results, simulation studies, and evaluation of out-of-sample performance. The theoretical results are scattered, and we summarize them. The results have been quite consistent: BMA has tended to outperform competing methods for model selection and taking account of model uncertainty. The theoretical results depend on the assumption that the "practical distribution" over which the performance of methods is assessed is the same as the prior distribution used, and we investigate sensitivity of results to this assumption in a simple normal example; they turn out not to be unduly sensitive.

We point out that HC's risk results, that AIC-model averaging and similar methods such as FIC-based model averaging perform well, depend crucially on their local misspecification assumption (2.2), namely that all nuisance parameters are small and decline with sample size, at rate $O(\frac{1}{\sqrt{n}})$. The key question is thus the realism of this assumption. We question this assumption on the grounds of its lack of face validity in some situations, the growing separation between data collection and research, the increasing tendency for research on different questions to be based on a few large high-quality public datasets, and the statistical literature, where sample size and parameter values rarely covary in the design of simulation studies. Finally, we reanalyze HC's data example, on risk factors for low birthweight.

Contents

List of Figures

List of Tables

1 Introduction

In their article, "Frequentist Model Average Estimators," Hjort and Claeskens — hereafter HC — make the point that statistical inference conditional on a model selected among several on the basis of data will tend to underestimate variability. We strongly agree. They argue that the way to overcome this is by model averaging, and again we agree. There is much support for these arguments: these points have been made by many authors in a long line of literature going back at least to Leamer (1977). HC point out that Bayesian model averaging (BMA) dominates the literature on accounting for model uncertainty in statistical inference. Their search for a frequentist alternative is largely motivated by the feeling that the performance of BMA in repeated datasets or experiments has been inadequately studied. Or, as they put it, "even though BMA 'works',..., rather little appears to be known about the actual performance or behavior of the consequent inferences, like estimator precision."

This is a somewhat surprising statement, as the performance of Bayesian model selection and BMA has, in fact, been extensively studied. There are three main strands of results: general theoretical results going back to Jeffreys (1939), simulation studies, and results on out-of-sample predictive performance. HC do not refer to any of this literature. The theoretical results are well-known but somewhat scattered in the literature. In brief, when used for model selection, the Bayes factor minimizes the Total Error Rate (sum of Type I and Type II error probabilities); BMA point estimators and predictions minimize mean squared error; BMA estimation and prediction intervals are calibrated; and BMA predictive distributions have optimal performance in the log score sense. We bring these results together in our Section 2. These results for BMA are quite general, and do not rely on the assumption that all uncertain parameters are small (essentially HC's local misspecification assumption, required by FMA). They also do not require the standard regularity conditions assumed by HC in deriving FMA, which are violated in many models of practical interest, such as change point models, or models involving unknown population size.

There are also several realistic simulation studies of the performance of BMA relative to other methods in a variety of situations, including linear regression (George and McCulloch 1993; Raftery, Madigan, and Hoeting 1997), loglinear models (Clyde 1999), logistic regression (Viallefont, Raftery, and Richardson 2001), and wavelets (Clyde and George 2000). In these studies, BMA was compared to the prevailing state of the art methods, and generally found to have better performance.

Finally, there has been extensive investigation of the out-of-sample predictive perfor-

mance of BMA compared to other methods, for real datasets. This is particularly important because these are situations in which the model assumptions underlying BMA and other methods do not necessarily hold, and they provide a neutral criterion for comparing methods. These include graphical models (Madigan and Raftery 1994; Madigan, Gavrin, and Raftery 1995), survival analysis (Raftery, Madigan, and Volinsky 1995), linear regression (Raftery, Madigan, and Hoeting 1997; Hoeting, Madigan, Raftery, and Volinsky 1999; Fernández, Ley, and Steel 2001a; Fernández, Ley, and Steel 2001b; Hoeting, Raftery, and Madigan 2002), binary regression (Fernández, Ley, and Steel 2002), and semiparametric regression (Lamon and Clyde 2000). The results of these studies have been quite consistent: BMA had better predictive performance than competing methods. It would be interesting to assess the predictive performance of FMA in the same way, using out-of-sample predictive performance. As HC note in their Section 10.5, the only model averaging methods that they discuss that have optimality properties are the Bayesian ones; FMA itself does not appear to yield optimal methods.

FMA consists of the analysis of the long-run properties of model averaging schemes *under the local misspecification assumption in HC's equation (2.2)*, which is essentially an assumption that all the parameters of interest for model averaging are small, specifically $O(\frac{1}{\sqrt{n}})$, modulo the known shift γ_0. As we discuss in our Section 4, this assumption is highly consequential, and HC's risk results depend on it crucially. As such, its realism is a critical issue, and we discuss that in Section 4.

HC's analysis of BMA under this local assumption, and their local approximation to Bayes factors, BLIC, are interesting and potentially relevant if one does accept the assumption. These ideas have been discussed previously. Smith and Spiegelhalter (1980) analyzed Bayesian model selection under similar local assumptions, and proposed several "local Bayes factors," derived in essentially the same way as HC's BLIC. It would be interesting to compare the two.

HC's proposal to estimate the spread in the BMA prior using empirical Bayes methods also seems useful. However, this idea has also been discussed previously, and in more depth. Volinsky (1997) suggested combining ridge regression and BMA in, essentially, an empirical Bayes BMA scheme. George and Foster (2000) proposed using the data to estimate the prior spread, and also the prior probability of a parameter being nonzero, yielding an empirical Bayes Bayesian variable selection method; this was extended to model averaging by Clyde and George (2000) in the context of wavelets. Hansen and Yu (2001) developed local empirical Bayes approaches in which the prior depends on the model in a data-dependent way, as

in HC's BLIC*.

This comment is organized as follows. In Section 2 we summarize some of the theoretical results in the literature on the performance of Bayesian model selection and BMA. These rely on the assumption that the prior distribution is representative of situations encountered in practice, and in Section 3 we investigate robustness to this assumption in a simple situation; the results seem fairly insensitive to this assumption. In Section 4 we show that the local misspecification assumption is important for HC's risk results, and we discuss its realism. In Section 5 we discuss the logistic regression example which is HC's only data example.

2 Performance of Bayesian Model Selection and Bayesian Model Averaging: Theoretical Results

Our goal is to make statements about the long-run performance of model averaging and the associated estimators. But which long run? In general, the performance of statistical methods depends on the underlying state of nature; there is no method which is uniformly optimal. Exceptions to this arise in special cases, for example in some estimation problems when a pivotal quantity is available. Thus we seek good performance on average over a range of situations, for example over the statistician's "career" of working with the model class in question. This involves averaging over situations where the different models hold (at least approximately). Within a given model, it involves averaging over a range or distribution of parameters typical of those encountered in practice. We call this the *practical distribution* of the parameters. This idea goes back at least to Jeffreys (1939), who referred to it using the term "world frequencies".

The first key result is due to Jeffreys (1939, p.327), , and concerns testing for two nested models.

Theorem 1 (Jeffreys) *For two nested models, model choice based on the Bayes factor minimizes the Total Error Rate (= Type I Error Rate + Type II Error Rate), if the practical distribution is equal to the prior distribution over the parameters.*

This leads to a different interpretation of the prior distribution from the usual one, namely the distribution of the parameters over which we would like to see good performance of the model choice method. Frequentist research on testing methods routinely implicitly defines such a distribution through the parameter values chosen for the simulation studies carried out to assess the power of a test; for recent articles of this kind in this journal, see Pena

and Rodríguez (2002) and Horowitz and Spokoiny (2002). Note that Theorem 1 generalizes immediately to the situation where the costs of the two types of error are unequal, by multiplying the Bayes factor by the ratio of the costs.

We now consider point estimation and point prediction. The BMA posterior distribution of a quantity of interest Q, which may be a model parameter, a "focus parameter" in HC's terminology, or an observable quantity to be predicted, is

$$p(Q|\text{data}) = \sum_S p(Q|S, \text{data})p(S|\text{data}), \tag{1}$$

where $p(Q|S, \text{data})$ is the posterior distribution of Q under model S and $p(S|\text{data})$ is the posterior probability of model S. It follows that the BMA point estimate of Q is

$$\hat{Q}_{\text{BMA}} = \sum_S \hat{Q}_S p(S|\text{data}). \tag{2}$$

This is of the type of HC's equation (4.1). As HC point out, for this to be valid, Q must have the same interpretation under all the models. What precisely this means has not been spelled out, as far as we know. We suggest one meaning: that Q be interpretable as a quantity that could be calculated from future data, at least asymptotically. The following result was alluded to in HC's Section 10.5:

Theorem 2 *$\hat{Q}_{\text{BMA}}$ minimizes MSE among point estimators, when the practical distribution of the parameters is equal to the prior distribution.*

We now consider interval estimation. We consider BMA estimation intervals with posterior content α. Then we have

Theorem 3 *BMA estimation intervals are calibrated, in the sense that the average coverage probability of a BMA interval with posterior content α is greater than or equal to α, on average over datasets drawn from the practical distribution, if the practical distribution of the parameters is equal to the prior distribution.*

Note that the BMA distribution of a quantity of interest can be viewed as the posterior distribution from the full model, with a mixed discrete-continuous prior distribution that assigns weight to the events that the individual components of γ are zero. Then Theorem 3 follows from the arguments of Rubin and Schenker (1986), with the continuous prior distribution used there replaced by the mixed discrete-continuous prior measure induced by

BMA. The only reason that the average coverage probability of the BMA interval is not exactly equal to α (rather than greater than or equal to α) is that for some datasets the interval may consist of just a single value corresponding to a component of γ_0 with posterior probability greater than α. Also, if the BMA estimation intervals are the shortest intervals with posterior content α, then they are the shortest intervals with the calibration property of Theorem 3.

Finally, we consider prediction of an observable out-of-sample quantity. Theorems 2 and 3 already show that BMA point prediction minimizes predictive MSE and that BMA prediction intervals are calibrated. The following further optimality result for BMA prediction was given by Madigan and Raftery (1994).

Theorem 4 *The BMA predictive distribution of a predictand Q is optimal under Good (1952)'s logarithmic scoring criterion:*

$$E\left[\log\left\{\sum_S p(Q|S, data)p(S|data)\right\}\right] \geq E[\log g(Q|data)]$$

for any probability distribution $g(\cdot|data)$, where the expectation is with respect to the predictive distribution $\sum_S p(Q|S, data)p(S|data)$.

This follows from the nonnegativity of the Kullback-Leibler information divergence. One way of interpreting this result is in terms of a simulation experiment. Data are generated from a model chosen at random among those considered, with parameters chosen at random from the prior distribution. The predictand Q is generated from the same model and parameter values, but these are unknown to the person forming the predictive distribution. Predictive distributions are generated using BMA, and any other competing method considered. Theorem 4 says that the log score will be better for BMA than for any other way of forming predictive distributions.

The results in this section are impressive, but they leave two questions open. How much better is BMA than other methods in specific situations? And how robust are these optimality results to the assumption that the prior distribution is equal to the practical distribution? We consider these questions in the context of a very simple example in the next section.

3 Normal Example

To get a sense of numerical differences in performance, and also of the extent to which the results in the last section hold even if the prior distribution is not the same as the

practical distribution, we consider a very simple normal example. Here data $(y_1, \ldots, y_n)$ are iid $N(\mu, 1)$ and we consider just two models, $M_0 : \mu = 0$ and $M_1 : \mu \neq 0$. Under M_1, the prior distribution is $\mu \sim N(0, \sigma^2)$. (The choice $\sigma^2 = 1$ is the Unit Information Prior and is generally agreed to be conservative (Raftery 1999), so here we consider only $\sigma^2 \leq 1$.) The practical distribution we consider draws equally from the two models, and under M_1, μ comes from a $N(0, \tau^2)$ distribution. The prior distribution is equal to the practical distribution when $\sigma = \tau$.

Analytic results are available in this situation. Although a simple special case, it is more general than it seems, because the results carry over fairly directly to one-degree-of-freedom nested model comparisons with at least moderate sample sizes under standard regularity conditions.

Figure 1 shows the total error rate when $n = 100$. The results were similar over a wide range of values of the practical variance τ^2, from $\frac{1}{16}$ to 1, so we show only the results for $\tau^2 = 1$. Model choice using a Bayes factor depends on the prior variance σ^2, and has a lower total error rate than a 5% significance test for all $\sigma^2 > 0.4$, i.e. for a prior variance that is "misspecified" relative to the practical distribution by a factor of up to 2.5. The Bayes factor has a lower error rate than AIC for all $\sigma^2 > 0.05$, i.e. for prior variances that are misspecified by a factor of up to 20. It would be interesting to extend this comparison to include FIC. Section 5.6 of Claeskens and Hjort, "The Focussed Information Criterion," suggests that FIC is likely to be close to AIC in practice, so one may conjecture that its performance will also be close to that of AIC.

Figure 2 shows the total error rate with a much larger sample size, $n = 100,000$. There we see that Bayes factors (and BIC) have much lower error rates than other methods, for all values of the prior variance. Once again, the result is robust to the practical variance τ^2, at least up to the factor of 16 in our experiments.

We now turn to estimation of μ. Figure 3 shows the mean squared errors of the BMA estimator with $n = 100$. These are compared with the MSE of the usual estimator $\hat{\mu} = \bar{y}$, which is $\frac{1}{n} = 0.01$. The BMA estimator outperformed the usual estimator by about 28% in terms of MSE, as long as the prior variance was not unduly small — greater than about 0.25. As can be seen from Figure 3, this result is robust to both the prior variance and the practical variance, and holds even when they differ by a factor of up to 4.

We now consider the coverage of interval estimates of μ. These are shown in Figure 4 for the BMA interval estimate and the standard normal confidence interval. The BMA interval has nearly correct coverage as long as the prior variance is not too small — again at least

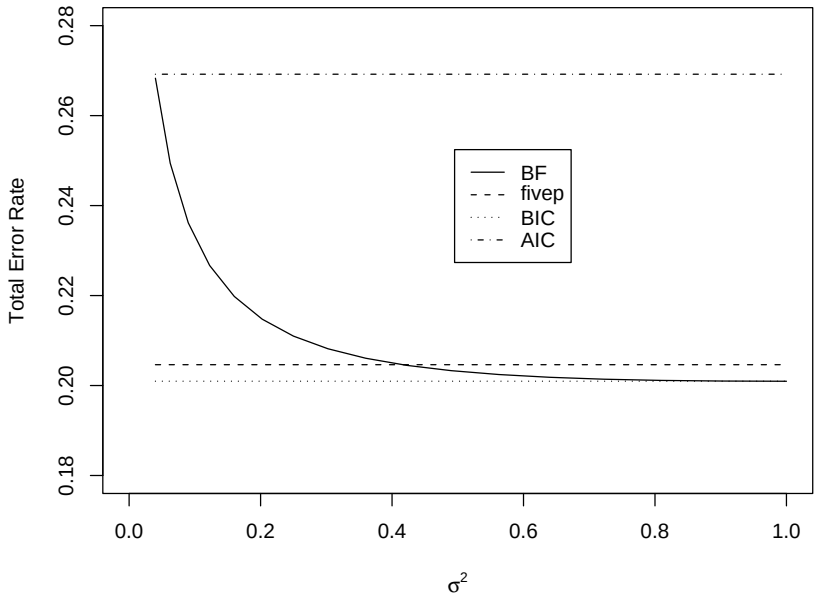

Figure 1: Total Error Rate in the Simple Normal Example for $n = 100$. Model choice is based on a Bayes Factor (solid line), a 5% significance test (dashes), BIC (dots), and AIC (dots and dashes). The x-axis shows the prior variance σ^2.

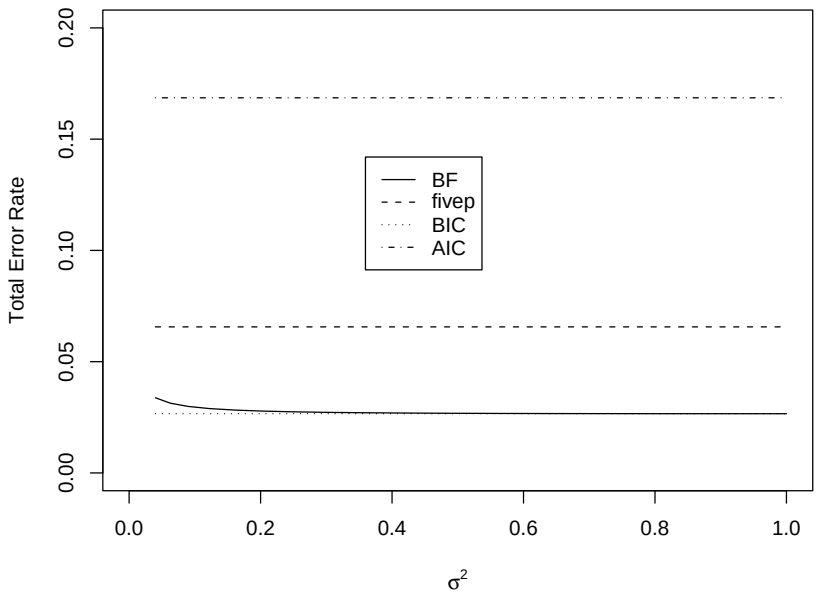

Figure 2: Total Error Rate in the Simple Normal Example for $n = 100,000$.

0.25.

The average length of the confidence intervals is shown in Figure 5. The BMA interval is consistently shorter than the standard confidence interval, by amounts that depend on the practical variance but are relatively insensitive to the prior variance. For unit prior variance, the reductions in the length of the confidence intervals range from 6% to 40%. It would be of interest to do a similar calculation for the AIC-based and FIC-based model averaging estimators.

4 The Local Misspecification Assumption, AIC and FMA

The results in Section 2 say that BMA is optimal in terms of mean squared error, and yields calibrated interval estimators of minimal length, provided that the prior distribution is equal to the practical distribution over which performance is assessed. HC assume that the practical distribution has variance $O(\frac{1}{n})$ — this is the local misspecification assumption in their (2.2). It follows from the arguments of Akaike (1983)that AIC provides an asymptotic ap-

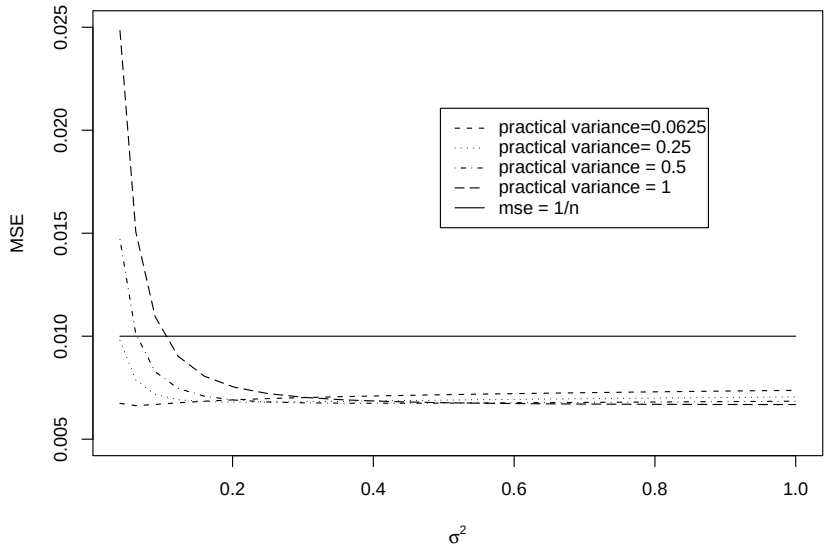

Figure 3: BMA Estimation of μ in the Simple Normal Example: Mean Squared Errors. The solid line shows the MSE for the standard estimator $\hat{\mu} = \bar{y}$, which is $1/n = .01$.

proximation to (twice the logarithm of) the Bayes factor provided that the prior distribution of the parameters contains about the same amount of information as the data, implying that the prior variance is $O(\frac{1}{n})$.

In the simple normal example of the last section, it can be shown that if the prior variance is proportional to a power of n, i.e. if $\sigma^2 = cn^{-\delta}$, then $\text{AIC} - 2\log B_{10} = O(1)$ if and only if $\delta = 1$, where B_{10} is the Bayes factor for M_1 against M_0; otherwise, if $0 \leq \delta < 1$, $\text{AIC} - 2\log B_{10} > O(1)$. Further, AIC is an unbiased estimator of twice the log Bayes factor under M_1, i.e.

$$E[\text{AIC} - 2\log B_{10}|M_1] = 0,$$

if and only if $c = e - 1 = 1.718$. Thus, in this case, AIC is equivalent to a Bayes factor if the prior contains the same information as about $0.58n$ observations. This seems like an unreasonably informative prior for many purposes.

In light of this, HC's risk results, that model averaging with AIC-like weights do well, are not surprising. In their case, both the prior and practical distributions have variances $O(\frac{1}{n})$. In this situation, the results in our Section 2 suggest that AIC-based model averaging will be close to optimal, and that BIC and standard Bayes factors will not, because the

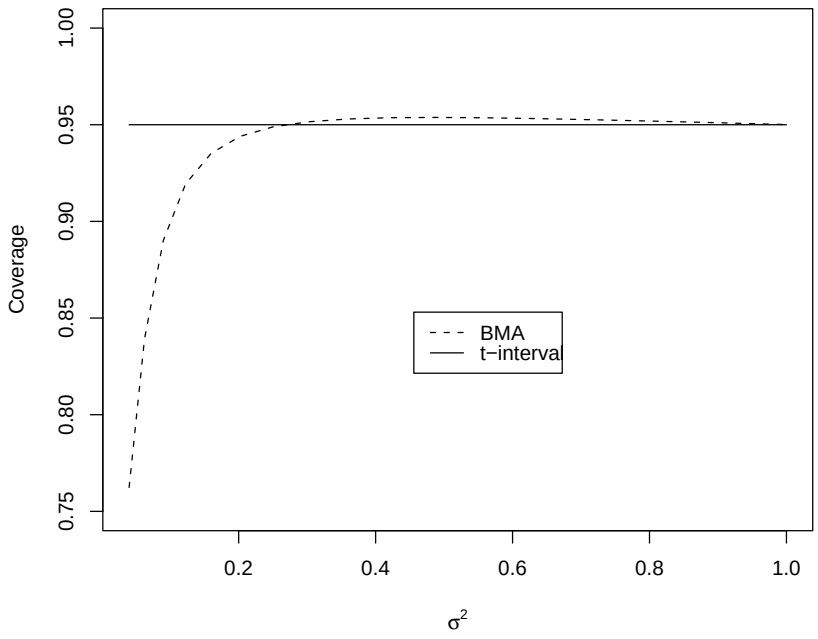

Figure 4: Coverage of 95% Confidence Intervals for μ in the Simple Normal Example: (a) BMA interval, and (b) standard normal confidence interval.

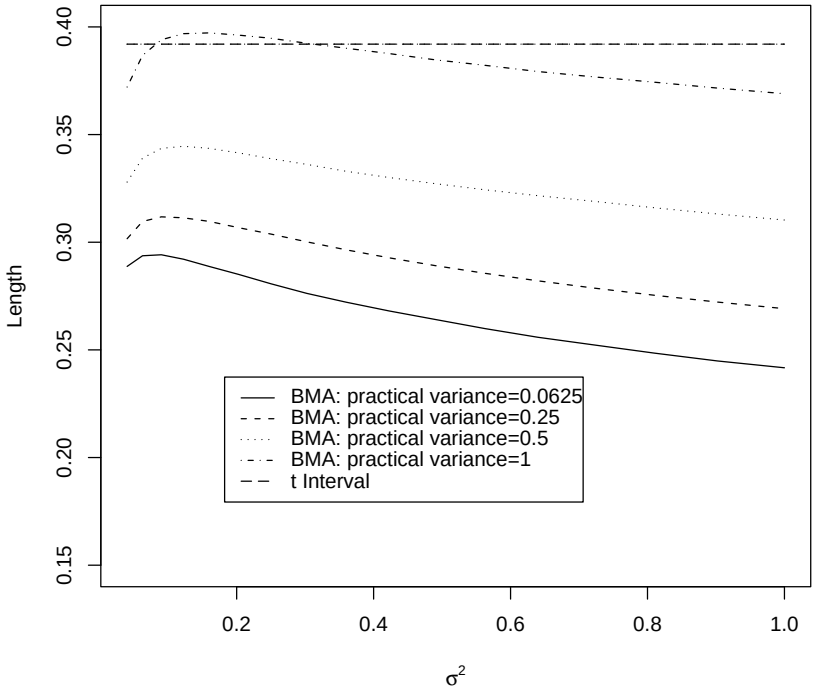

Figure 5: Average Lengths of Confidence Intervals for μ in the Simple Normal Example

prior on which they are based is very different from the practical distribution used to assess them. FIC-based model averaging is a further refinement designed specifically to give optimal estimation results *when the local misspecification assumption holds*, and so it is no surprise that it does well in the simulation studies designed by the authors.

Thus the local misspecification assumption (2.2) is critical to HC's results. It is not just a technical regularity condition, but a key assumption about the way the world works. The question then becomes, it is realistic? There are two standard arguments for its realism. One is that as the same size increases, the effects or parameters of interest become smaller — in essence, give researchers a more powerful microscope, and they will look for smaller objects. The other is the converse, that researchers aiming at finding small effects are more likely to use larger sample sizes. These arguments support the direction of the association between parameter size and sample size, but not the assumed rate.

We are not convinced, however. The local misspecification assumption applies to *nuisance parameters* rather than to quantities of primary interest, and the main arguments supporting the assumption refer mainly to quantities of primary interest. It does not seem to cover the situation where individual γ_i are not small, but model uncertainty is still present because of correlation between the corresponding x_i. It also does not represent the situation where the coefficients for some nuisance variables are substantial, and those for others are small; in our experience this is a common situation.

The assumption seems implausible on its face in some situations. For example, consider the low birthweight example. Risk factors for low birthweight have been considered in large studies with thousands of subjects, as well as the small study with 189 subjects discussed by HC. The main nuisance parameter in HC's version of the problem is the effect of maternal age — is it reasonable to expect this to be much larger in the study discussed by HC than in another much larger one, just because the sample size is much smaller? It seems unlikely to us.

The salience of the local misspecification assumption seems likely to be diminishing given current trends in science. Increasingly, data collection and research are disassociated, with disciplines more and more organized around large, high-quality publicly available databases collected using public funds, and many researchers addressing different questions using the same data. This makes it less likely that the size of parameters would depend on sample size — researchers are more likely to choose datasets based on whether or not they contain variables of interest than on their sample size.

Sociology provides one example of this — much sociological research on very diverse

topics is carried out using a small number of large databases such as the General Social Survey (GSS), the National Longitudinal Survey of Youth (NLSY), and the National Survey of Families and Households (NSFH). These databases all have comparable sample sizes, on the order of 5,000 to 20,000. They are used to investigate all sorts of sociological questions and to estimate a wide range of parameters, large and small. The size of the effects of main interest is hardly related to the sample size (which is essentially constant and out of the researcher's control), and the size of the nuisance parameters is even less likely to be related to sample size.

Another such discipline is astronomy, which is moving towards the same model as sociology, with many researchers working at their computer screens in a "virtual observatory" such as Skyview (`http://skyview.gsfc.nasa.gov`), rather than generating their own data. Epidemiology — long a bastion of research-group-specific datasets — is also moving in this direction, albeit in a different way, via meta-analysis, with the pooling of all data from all available studies. A further example is political science — a great deal of North American political science research is based on a single database, the National Election Studies (NES — `http://www.umich.edu/ nes`). A bibliography lists roughly 4,000 publications based on the NES data.

When sample size is decided on by researchers in terms of the question being studied, it is often determined, not by the size of the effect being studied, but by its *importance*. Large effects may actually be the object of *larger* studies. For example, the association between smoking and lung cancer is certainly a large one, and once its existence was suspected, several very large studies were carried out to assess it.

Do statisticians act as if they believe the local misspecification assumption? One way of assessing this, implicitly, is by looking at the design of simulation studies in the statistical literature that assess the performance of estimators and tests. If they did, one would expect to see simulation studies with a relationship between sample size and parameter value, with large sample sizes corresponding to small parameter values. This rarely, if ever, happens. Some examples from recent issues of JASA, are Horowitz and Spokoiny (2002), Pena and Rodríguez (2002) and Chaterjee, Chen, and Breslow (2003) — in their simulation studies, as in most others of which we are aware, parameter values and sample size were varied independently. It seems that statisticians do not see an inverse relationship between parameter size and sample size of the kind implied by the local misspecification assumption as an important enough feature of reality to be worth including in simulation studies.

5 Model Averaging for Logistic Regression

HC's only data example is the logistic regression for predicting low birth weight. Their "focus parameters" are the probability of low birth weight for a white mother with covariates equal to the average for whites in the study, the same quantity for a black mother, and the ratio of the two. The latter seems like a strange choice. If the ratio is different from one, this could be due to interracial differences in the probability of low birth weight, in the average covariates, or both; the measure conflates the two sources of variation. In epidemiological studies, interest generally focuses on the extent to which an independent variable of interest (here race) is a risk factor, after adjusting for other covariates — in the present context this is just the logistic regression parameter for black (x_4). Epidemiologists are also interested in subpopulation average prevalences. However, the ratio focus parameter used by HC corresponds to neither of these, and it does not seem to provide an answer to any scientific question of wide interest.

5.1 Bayesian Model Averaging for Case-Control Studies

HC's analysis does not tell us how accurate any of the estimators or standard errors are in this example. It therefore seems to be of interest to summarize the only study that we know of the performance of model averaging for logistic regression (Viallefont, Raftery, and Richardson 2001). This was carried out in the context of what is probably the largest area of application of logistic regression: epidemiological case-control studies. Typically there is one "focus parameter" of interest — the adjusted effect of a potential risk factor of interest, as measured by the logistic regression parameter. Usually there are many potential confounders, on the order of dozens, and the task is to make inference about the effect of the risk factor of interest.

BMA was implemented for this application using a prior distribution for the effect of interest that was agreed by a team of collaborating epidemiologists, and that implied that the odds ratio was unlikely to be greater than 7. Model averaging was carried out using the `glib` software for BMA in generalized linear models (Raftery 1996), available at `www.research.att.com/~volinsky/bma.html`. The performance of BMA and other confounder selection methods was analyzed by means of a simulation study whose specification (numbers of cases and controls, numbers and effect sizes of potential confounders, actual odds-ratios) was based on a sample of 50 case-control studies in the epidemiological literature. It is often possible to design a simulation study to favor almost any model selection

or averaging method, and basing the design on a sample of actual studies helps to minimize such biases.

The results were as follows. The BMA posterior probability of the adjusted odds ratio of interest being different from 1, averaged over all models, was well calibrated, while significance tests with standard confounder selection methods were not. BMA interval estimates were well calibrated, and BMA point estimates had MSE about 20% lower than standard variable selection methods.

5.2 Bayesian Model Averaging for the Low Birthweight Example

We now give BMA results for the low birthweight example. As we have noted, HC's main focus parameter seems of dubious scientific value, but we give results for it anyhow. Also, HC have greatly simplified the model uncertainty aspect of the problem. In the initial dataset of Hosmer and Lemeshow (1989), there were nine independent variables about which there was uncertainty (counting the two race dummy variables). However, HC removed five of the variables from the dataset, namely smoking, history of premature labor, history of hypertension, uterine irritability, and number of physician visits. They also assumed that there is no uncertainty about the inclusion of the maternal weight variable, thus reducing the number of uncertain variables from nine to three, and the number of potential models from 528 to 8. First we give BMA results on the same basis as the HC analysis, and in Section 5.3 we give BMA results for the complete problem.

We compute posterior model probabilities in four ways. First, we use the reference proper prior approach of Raftery (1996) with prior dispersion parameter $\phi = 1$. While proper, this prior is designed to be spread out enough as to be essentially noninformative; the prior standard deviation of the "black" effect, the regression parameter for x_4, is 6.3. Weakliem (1999) has argued that odds ratios greater than about 15 are unusual in social scientific contexts of this kind, and we translate that into an "informative" prior for the "black" and "other race" parameters that has standard deviation 1.35. We compare these with model averaging using the BIC approximation and the AIC weights.

Table 1 shows the standard frequentist results and the BMA posterior model probabilities for the 8 models considered by HC. None of the larger models fits significantly better than the "narrow" model by standard criteria at the 5% level, and the reference BMA analysis as well as the BIC approximation favor the narrow model, although not decisively, agreeing with the standard analysis. The BMA analysis with an informative prior gives more weight to the wider models; this is due to the additional information in the prior. Model averaging with

Table 1: Standard GLIM Analysis and Posterior Model Probabilities for HC's Subset of the Low Birthweight Data

Model	Dev diff	df	P value	Posterior Model Probabilities (%) Reference Prior	Informative Prior	BIC Approx	AIC Weights
0	0	0	—	54	25	54	11
3	1.57	1	.21	8	4	9	9
4	3.62	1	.06	24	38	24	24
5	0.59	1	.44	5	7	5	5
34	4.52	2	.10	3	4	3	14
35	2.01	2	.37	1	1	1	4
45	5.43	2	.07	5	20	4	22
345	6.03	3	.11	0	2	0	11

NOTE: Dev diff is the deviance difference between the model considered and HC's "narrow" model with just maternal weight as covariate.
df refers to the number of degrees of freedom in the comparison, and P value to the P value for the asymptotic χ^2 distribution of the deviance in testing the model considered against the narrow model.

AIC weights also gives more weight to the wider models; this can be viewed as a consequence of the fact that this is a form of BMA with quite informative prior distributions. The BIC and reference BMA analyses are in close agreement, which is to be expected as both correspond to the use of a unit information prior for the parameters (Kass and Wasserman 1995; Raftery 1995, 1996). .

Table 2 shows the BMA estimators and posterior standard deviations for HC's focus parameters, and may be compared with the table in HC's Section 6.2. The results are fairly similar across model averaging methods. The difference between model selection and model averaging is especially striking for the reference prior BMA and the BIC approximation, which favor the narrow model. For the narrow model, the standard error of the ratio focus parameter is 0.06, while for BMA it is 0.42.

5.3 Analysis of Complete Low Birthweight Data

HC excluded the five variables smoking, premature labor, hypertension, uterine irritability, and physician visits from the analysis, but did not discuss this decision; we could not see that it would lead to better inferences, whether one is interested in the association between race

Table 2: BMA Estimates and Posterior Standard Deviations for HC's Focus Parameters for HC's Subset of the Low Birthweight Data

	Reference Prior	Informative Prior	BIC Approx	AIC Weights
For p(white):				
estimate	.285	.268	.285	.261
stdev	.040	.045	.040	.046
For p(black):				
estimate	.306	.357	.306	.369
stdev	.098	.113	.098	.112
For the ratio:				
estimate	1.096	1.359	1.094	1.442
stdev	.420	.532	.418	.549

and low birthweight after adjusting for other factors, or in explaining the total association between race and low birthweight in terms of other factors. Also, we were unclear about the justification for HC's decision to include maternal weight with prior probability 1.0. Hosmer and Lemeshow (1989) themselves made inference about this from the data at hand rather than *a priori*: the purpose of their study was to find out which of the collected variables, all known to be associated with low birthweight in some populations, were important in the population being served by the medical center where the data were collected; see Hosmer and Lemeshow (1989, pp. 91–94). As already mentioned, we were also unclear about the choice of focus parameters, which seem to differ from standard epidemiological practice.

As a result, we reanalyzed the dataset, including all the variables and taking account of uncertainty about them, with a focus on the logistic regression parameters themselves, which correspond to adjusted log-odds ratios. This leads us to 528 models rather than HC's 8. We first carried out a reference prior BMA analysis (Raftery 1996); as before, the results for this were similar to those using BMA with the BIC approximation. We then carried out a more informative analysis using the prior with standard deviation 1.35 for the last seven variables, all of which are either binary or counts. For computational convenience, we excluded the models whose BIC-approximated posterior probability was less than that of the most likely model by a factor of 20 or more; this step is optional and the results are insensitive to it. This left 86 models that we averaged over.

This analysis can be done easily in Splus using the two commands:

Table 3: Posterior Effect Probabilities, BMA Posterior Means, and BMA Posterior Standard Deviations for the Full Low Birthweight Dataset

Parameter	Reference Prior			Informative Prior		
	$\Pr[\beta \neq 0]$	Mean	SD	$\Pr[\beta \neq 0]$	Mean	SD
Age	8	.048	.035	3	−.046	.035
Maternal weight	71	−.016	.007	72	−.016	.007
Black	25	.986	.509	58	.910	.478
Other race	17	.750	.466	47	.717	.417
Smoking	36	.772	.391	68	.771	.381
Premature labor	42	.719	.335	46	.627	.328
Hypertension	68	1.761	.713	88	1.352	.623
Uterine irritability	29	.886	.443	49	.780	.423
Physician visits	1	−.059	.168	1	−.064	.167

NOTE: These results are based on Bayesian model averaging across the 86 models whose BIC-approximated posterior probabilities were at least 1/20 of that of the model with the highest one.
$\Pr[\beta \neq 0]$ is the posterior effect probability, i.e. the probability, given the data, that the parameter is different from zero, expressed as a percentage.
The posterior mean and standard deviation are calculated conditionally on the variable being in the model, i.e. on the associated regression parameter being different from zero.

```
bic.hosmer <- bic.glm (x,y,binomial)
glib.hosmer <- glib (x,y,error="binomial",link="logit",models=(bic.hosmer$which)*1,phi=1)
```

where `bic.glm` and `glib` are Splus functions that can be downloaded from `www.research.att.com/~volinsky/bma.html`, `x` is the 189×9 design matrix of independent variables, and `y` is the vector of responses. The BMA analysis with informative priors requires specification of the `priorvar` matrix argument in `glib`. The results are shown in Table 3.

The posterior effect probabilities in Table 3 can be interpreted in light of the commonly used scale for Bayes factors (Jeffreys 1939; Kass and Raftery 1995), on which odds of less than 3:1 are viewed as weak evidence. Thus posterior effect probabilities between 25% and 75% would correspond to weak evidence one way or the other. Most of the effects in this dataset are within this indecisive range. The additional information in the informative prior tends to increase the evidence for individual parameters, but generally not enough to change the qualitative conclusion. The most likely single model includes all the variables except age, premature labor and physician visits. In most cases, the posterior effect probabilities reflect more uncertainty than P values based on a single model; this is due in part to taking

account of model uncertainty.

References

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute 44*, 277–291.

Chaterjee, N., Y. H. Chen, and N. E. Breslow (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association 98*, 158–168.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies (with Discussion). In *Bayesian Statistics 6* (edited by J. M. Bernardo et al.), pp. 157–185. Oxford, U.K.: Oxford University Press.

Clyde, M. A. and E. I. George (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, series B 62*, 681–698.

Fernández, C., E. Ley, and M. F. J. Steel (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics 100*, 381–427.

Fernández, C., E. Ley, and M. F. J. Steel (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*, 563–576.

Fernández, C., E. Ley, and M. F. J. Steel (2002). Bayesian modelling of catch in a northwest Atlantic fishery. *Applied Statistics 51*, 257–280.

George, E. I. and D. P. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika 87*, 731–747.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*, 881–889.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, series B 14*, 107–114.

Hansen, M. H. and B. Yu (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association 96*, 746–774.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science 14*, 382–417. [A corrected version is available at `www.stat.washington.edu/www/research/online/hoeting1999.pdf`.].

Hoeting, J. A., A. E. Raftery, and D. Madigan (2002). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics 11*, 485–507.

Horowitz, J. L. and V. G. Spokoiny (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association 97*, 822–835.

Hosmer, D. W. and S. Lemeshow (1989). *Applied Logistic Regression*. New York: Wiley.

Jeffreys, H. (1939). *Theory of Probability*. Oxford, U. K. : Oxford University Press.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*, 928–934.

Lamon, E. C. and M. A. Clyde (2000). Accounting for model uncertainty in prediction of chlorophyll a in Lake Okeechobee. *Journal of Agricultural, Biological, and Environmental Statistics 5*, 297–322.

Leamer, E. E. (1977). *Specification Searches: Ad Hoc Inference With Nonexperimental Data*. New York: Wiley.

Madigan, D., J. Gavrin, and A. E. Raftery (1995). Enhancing the predictive performance of Bayesian graphical models. *Communications in Statistics - Theory and Methods 24*, 2271–2292.

Madigan, D. and A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association 89*, 1535–1546.

Pena, D. and J. Rodríguez (2002). A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association 97*, 601–610.

Raftery, A. E. (1995). Bayesian model selection in social research (with Discussion). In *Sociological Methodology 1995* (edited by P. V. Marsden), pp. 111–163. Cambridge, Mass. : Blackwell Publishers.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika 83*, 251–266.

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods and Research 27*, 411–427.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association 92*, 179–191.

Raftery, A. E., D. Madigan, and C. T. Volinsky (1995). Accounting for model uncertainty in survival analysis improves predictive performance (with Discussion). In *Bayesian Statistics 5* (J. M. Bernardo *et al.* , eds.), pp. 323–349. Oxford, U.K.: Oxford University Press.

Rubin, D. B. and N. Schenker (1986). Efficiently simulating the coverage properties of interval estimates. *Applied Statistics 35*, 159–167.

Smith, A. F. M. and D. J. Spiegelhalter (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, series B 42*, 213–220.

Viallefont, V., A. E. Raftery, and S. Richardson (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine 20*, 3215–3230.

Volinsky, C. T. (1997). *Bayesian Model Averaging for Censored Survival Models*. Ph. D. thesis, University of Washington, Seattle.

Weakliem, D. L. (1999). A critique of the Bayesian Information Criterion. *Sociological Methods and Research 27*, 359–397.